Application of Machine Learning Approaches on Quantitative Structure Activity Relationships

Mariusz Butkiewicz, Ralf Mueller, Danilo Selic, Eric Dawson, and Jens Meiler

***Abstract*—Machine Learning techniques are successfully applied to establish quantitative relations between chemical structure and biological activity (QSAR), i.e. classify compounds as active or inactive with respect to a specific target biological system. This paper presents a comparison of Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Decision Trees (DT) in an effort to identify potentiators of metabotropic glutamate receptor 5 (mGluR5), compounds that have potential as novel treatments against schizophrenia. When training and testing each of the three techniques on the same dataset enrichments of 61, 64, and 43 were obtained and an area under the curve (AUC) of 0.77, 0.78, and 0.63 was determined for ANNs, SVMs, and DTs, respectively. For the top percentile of predicted active compounds, the true positives for all three methods were highly similar, while the inactives were diverse offering the potential use of jury approaches to improve prediction accuracy.**

***Index Terms*—Machine Learning, quantitative structure activity relationship (QSAR), Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees (DT), area under the curve (AUC), receiver operator characteristics (ROC), high-throughput screening (HTS)**

I. Introduction

In this paper we present a comparison of three machine learning techniques applied to a specific Quantitative Structure Activity Relationship (QSAR) [1], [2] problem. Machine Learning algorithms have proven to be of practical value for approximating nonlinear separable data, especially for classifying biological target data [3], [4]. Artificial neural networks (ANN) [5], [6] support vector machines (SVM) [7], [8] as well as decision trees (DT) [9] have been applied in the past.

Burton e*t al.* [10] reviewed application of several types of ANNs for establishing QSARs and highlighted potential difficulties and challenges in their application. An overview of ANNs, their limitations, and their use in evaluating chemical structure data was presented by Winkler [5]. Hecht *et al.* [11] predict dihydrofolate reductase inhibition based on data derived from high-throughput screening (HTS). Fogel [12] analyzed a combination of clustering and ANNs prescreen compounds for HIV inhibition optimizing specificity and potency.

Fang *et al.* [13] presented an effective application of SVMs in mining HTS data from a type I methionine aminopeptidases (MetAPs) inhibition study. This method was applied on a compound library of 43,736 organic small molecules and 50% of the active molecules could be recovered by screening just 7% of the test set. According to Plewczynski *et al.* [14], a SVM was able to achieve classification rates of up to 100% in evaluating the activity of compounds with respect to specific targets. Their overall hit rate, however, was somewhat lower, 80%. Stahura and Bajorath [15] looked at several computational approaches, including SVMs, as a way to complement HTS.

An approach combining SVMs and recursive partitioning by DTs to predict the metabolic stability of compounds is described by Sakiyama *et al.* [16]. The same publication also discusses logistic regression and random forest approaches. Similarly, Baurin *et al.* [17] consider numerous statistical and computational techniques, including DTs, in their 2D-QSAR models for COX-2 inhibition based on the 193,447-compound NCI database.

Burton *et al.* [10] applied DTs in combination with a statistical learning method for predicting the CYP1A2 and CYP2D6 inhibition. CYP2D6 datasets provided eleven models with an accuracy of over 80%, while CYP1A2 datasets counted five high-accuracy models for HTS. The application of DTs in drug discovery is discussed by Rusinko *et al.* [18]. Their research focuses on a dataset with 1,650 monoamine oxidase inhibitors. Recently, Simmons *et al.* [19], [20] described an ensemble based DT model to virtually screen and prioritize compounds for acquisition.

In the present work three different approaches are applied to *in silico* screening for potentiators of metabotropic glutamate receptor subtype 5 (mGluR5). Selective potentiators of the metabotropic glutamate receptor subtype mGluR5 have exciting potential for development of novel treatment strategies for schizophrenia and other disorders that disrupt cognitive function [21]. The latest generation of selective mGluR5 potentiators is based on the lead compound CDPPB and features systemically active compounds with long half-lives that cross the blood-brain barrier [22].

The accuracy of each of the machine learning techniques is depicted by way of receiver operating characteristic (ROC) curves [23] and compared by area under the curve (AUC) as

Manuscript received November 12, 2008. This work was supported in part by grant R21-MH082254 from the National Institute of Mental Health.

Mariusz Butkiewicz, Ralf Mueller, and Danilo Selic work at the Center for Structural Biology at Vanderbilt University, Nashville TN, 37232. These authors contributed equally to this study.

Eric Dawson, PhD, is a research assistant professor at the Center for Structural Biology at Vanderbilt University, Nashville TN, 37232

Corresponding author: Jens Meiler, PhD, Center for Structural Biology, Vanderbilt University, Nashville, TN 37232.

978-1-4244-2756-7/09/$25.00 ©2009 IEEE

well as enrichment values. This paper is organized as follows: Section II provides an overview of the Machine Learning techniques used, section III describes the generated QSAR training data and its descriptors. Section IV introduces the methods and their implementations. Section V states the result achieved with the approach. Concluding remarks are given in Section VI.

II. Machine Learning Techniques

A. Artificial Neural Networks

ANNs model the human brain and its capability to recognize patterns. Therefore, in the simplest ANN *in silico* model systems interlink several layers of neurons by weighted connections w_{ij}. The input data x_i to the first layer are summed up according to their weights and modified by the activation function *K*:

$$f_j(x) = K(\sum_i x_i w_{ij}) \quad (1)$$

The output f_j then serves as input to the *j*-th neuron of the next layer (Fig. 1).

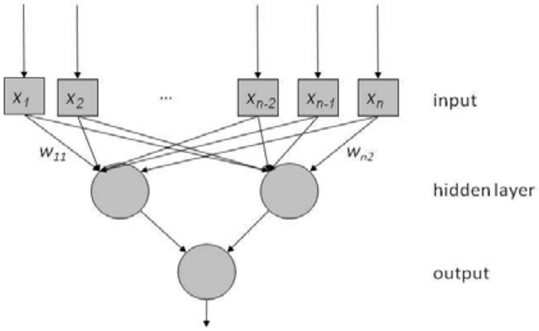

Fig. 1. Schematic view of an ANN: Schematic view of an ANN: Up to 1,252 descriptors are fed into the input layer. The weighted sum of the input data is modified by the activation function and serves as input to the next layer. The output describes the biological activity of the molecule.

The difference between the calculated output and the target value in a supervised training scheme determines the change of each weight (back-propagation of errors). The training of the ANN consists of an iteration of weight changes that minimizes the root mean square deviation *rmsd* between experimental and predicted biological activity. For an overview on ANNs and their application in chemistry see for instance Zupan [24].

$$rmsd = \sqrt{\frac{\sum_{i=1}^{n}(exp_i - pred_i)^2}{n}} \quad (2)$$

The present ANNs have up to 1,252 inputs, eight hidden neurons, and one output (biological activity). The logistic function

$$g(x) = \frac{1}{1+e^{-x}} \quad (3)$$

is applied as activation function *K* of the neurons. The training method used is Resilient Propagation [25], a supervised learning approach.

B. Support Vector Machines

SVM learning with the extension for regression estimation is the second approach of machine learning [26], [27]. The main characteristics are the estimation of the regression using linear functions defined in high-dimensional feature space [28], risk minimization according to Vapnik's ε - intensive loss function [29], as well as structural risk minimization which minimizes the risk function consisting of the empirical error and the regularized term [29].

A training data set can be described as $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$ with $i = 1, \dots, l$ where l is the total number of available input data pairs consisting of molecular descriptor data and biological activity. For the approximation the SVM considers the following function:

$$f(x, w) = w * \phi(x) + b \quad (4)$$

where $\phi(x)$ describes a function that performs a nonlinear transformation from the given input space X. The parameters w and b are estimated by finding the minimum of Vapnik's linear loss function with insensitivity zone as a measure of the error of approximation:

$$|y - f(x,w)| = \begin{cases} 0, if\ |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon, \quad otherwise \end{cases} \quad (5)$$

Thus, the loss is equal to zero if the difference between the predicted $f(x,w)$ and the measured value y is less than ε. Vapnik's insensitivity loss function defines an $\boldsymbol{\varepsilon}$ - tube (Fig 2).

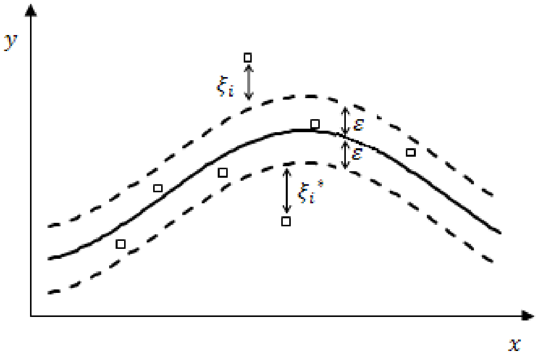

Fig. 2. Schematic of a Support Vector $\boldsymbol{\varepsilon}$ - tube : data points in $\boldsymbol{\varepsilon}$ – tube are not penalized, while points outside the tube get a penalty according to their distance from the tube edge.

If the predicted value is within the $\boldsymbol{\varepsilon}$ - tube the loss (error) is zero. For all other predicted points outside the tube, the loss equals the magnitude of the difference between the predicted value and the edge of the tube. Minimizing the following function L results in solving the regression problem:

$$L_{w,\xi,\xi^*} = \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^{\,*}) \qquad (6)$$

under constraints

$$\begin{aligned} & y_i - g(x,w) \leq \varepsilon + \xi_i \\ & g(x,w) - y_i \leq \varepsilon + \xi_i^{\,*} \\ & \xi_i^{(*)} \geq 0 \qquad\qquad i = 1, \dots, l \end{aligned} \qquad (7)$$

where ξ_i and $\xi_i^{\,*}$ are slack variables shown in Fig. 2 for measurements above and below an ε - tube, respectively. Both slack variables are positive values and their magnitude can be controlled by penalty parameter C. This optimization problem is then transformed into the dual problem, and its solution is given by:

$$\begin{aligned} & f(x) = \sum_{i=1}^{N_{SV}}(\alpha_i - \alpha_i^{\,*}) * K\left(x_{i,}x\right) + \ b \\ & 0 \leq \alpha_i \leq C, \\ & 0 \leq \alpha_i^{\,*} \leq C \end{aligned} \qquad (8)$$

where α_i and $\alpha_i^{\,*}$ are the Lagrange multipliers corresponding to ξ_i and $\xi_i^{\,*}$, N_{SV} is the number of support vectors SV and $K(x_i, x)$ is the kernel function. The Gaussian kernel was used to train the Support Vector Machine. The constant C influences a tradeoff between an approximation error and the weight vector $\|w\|$ norm. The optimal parameter C is chosen using cross validation on a monitoring dataset. An increase in C penalizes larger errors (large ξ_i and $\xi_i^{\,*}$) and leads to an decrease of the prediction error for the training dataset. However, this can be achieved only by increasing the weight vector norm $\|w\|$. While an increase in $\|w\|$ reduces the prediction error for the training dataset it does not guarantee a small generalization performance of a model due to possible over-fitting. Hence C needs to be optimized by minimizing the error of a monitoring dataset. Another design parameter is the required precision embodied in an $\boldsymbol{\varepsilon}$ value that defines the size of an $\boldsymbol{\varepsilon}$ - tube.

C. Decision Trees

The third type of machine learning approach used in this research is DT learning. Its output is a tree diagram or dendrogram (Fig. 3), a model that describes how a given dataset can be classified by assessing a number of predictor variables and a dependent variable.

$$(x, y) = (x_1, x_2, x_3 \dots, x_n, y) \qquad (9)$$

This is achieved by way of a partitioning algorithm, which gauges each predictor to determine which values of that predictor, if any, can be used to forecast the value of the dependant variable. The dataset is then successively split into subsets (nodes) by the descriptor that produces the greater purity in the resulting subsets. The predictive power of a descriptor can be ascertained in a number of ways.

For instance, the CHAID algorithm, developed by Kass [30] uses a chi-squared test for predictor evaluation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\mathrm{O_i} - \mathrm{E_i})^2}{\mathrm{E_i}} \qquad (10)$$

where χ^2 is the test statistic, $\mathrm{O_i}$ is the observed frequency, $\mathrm{E_i}$ is the expected frequency asserted by a null hypothesis and n the number of possible outcomes of each event.

Another common approach to selecting a descriptor for a split relies on entropy (11) and information gain (12) (e.g. ID3 and C4.5 algorithms [31]):

$$H(X) = \sum_{x \in X} p(x)\ log\ p(x) \qquad (11)$$

$$D_{KL} = \sum_{x \in X} p(x)\ log \frac{p(x)}{q(x)} \qquad (12)$$

In equation 11, *H(X)* denotes entropy, *X* is a discrete random variable, *x* a given state of *X*, and *p(x)* the probability of *x*. Equation 12 describes information gain, comparing a true probability distribution *p(X)* and an arbitrary probability distribution *q(X)*.

The splitting process continues until a predefined number of nodes has been created (or some other termination criterion is met) to avoid over-fitting. Alternatively, the entire tree can first be generated and then reduced in size by a technique called pruning that eliminates nodes and branches that are not statistically significant.

DT learning produces a sequence of splitting criteria, which after being established in an initial run (training), can be used to classify a new, independent dataset.

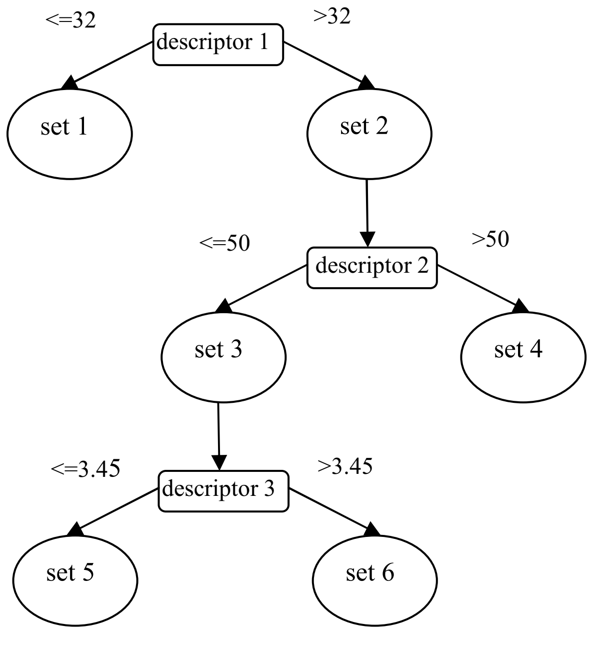

Fig. 3. An example of a decision tree. The initial dataset is first split according to *descriptor 1* into two subsets, *set 1* and *set 2*. *Set 2* is in turn partitioned into two more subsets by *descriptor2*, creating *set 3* and *set 4*. Finally, *set 3* is divided into *set 5* and *set 6* by *descriptor 3*.

III. Training Data

Glutamate is the primary excitatory neurotransmitter in the mammalian central nervous system (CNS) and is responsible for generation of fast excitatory synaptic responses at the vast majority of CNS synapses [32]. Fast synaptic responses at glutamatergic synapses are mediated by activation of a well characterized family of glutamate receptor cation channels referred to as the ionotropic glutamate receptors (iGluRs). In addition, glutamate activates metabotropic glutamate receptors (mGluRs), which are coupled to effector systems through GTP-binding proteins [33], [34]. The mGluRs provide a mechanism by which glutamate can modulate or fine tune activity at the same synapses at which it elicits fast synaptic responses. Because of the ubiquitous distribution of glutamatergic synapses, mGluRs participate in a wide variety of functions of the CNS [33], [35], [36].

A number of recent studies suggest that activation of one of the mGluRs, mGluR5, could have robust effects in forebrain circuits thought to be disrupted in schizophrenia. Hence, it was postulated that activators of mGluR5 could provide novel therapeutic agents that may be useful for treatment of this disorder [37], [38].

In a high throughput screen 144,475 compounds were tested for allosteric potentiation of mGluR5 using full automation in conjunction with the Vanderbilt HTS facility. Receptor-induced intracellular release of calcium was measured by utilizing an imaging-based plate reader that makes simultaneous measurements of calcium levels in each well of a 384 plate (Hamamatsu FDSS). Outliers were evaluated by visual inspection to ensure the quality of hit selection. Putative hits were confirmed and their concentration response was assayed. 1,387 compounds were verified as potentiators of mGluR5.

To apply the compound data towards the machines learning approaches, each molecule is numerically described by a molecular fingerprint. For the descriptor calculation (Table I) the external software suite ADRIANA [39] was utilized.

IV. Implementation / Method

The BioChemistryLibrary (BCL) is a class library written in the C++ programming language that includes classes to model both rather small organic molecules and larger molecules such as proteins, DNA, and RNA. Both the ANNs and SVMs were implemented within this framework.

A third-party DT generation application called FIRM (Formal Inference-based Recursive Modeling) [40], which works with both categorical and continuous dependent variables and improves on the CHAID recursive partitioning algorithm [30], was used for the DT evaluation. FIRM was applied to the same training and independent datasets as the ANN and SVM.

A. Dataset generation

The datasets used in this research were derived from a database of 144,475 compounds as a maximally diverse subset of the commercially available compounds contained in the ChemBridge and ChemDiv libraries. An initial HTS revealed that 1,356 of these compounds were mGluR5 potentiators. Of the total, 14,448 (10%) compounds were set aside for monitoring and an additional 14,448 (10%) were reserved for independent testing of QSAR models, leaving 115,581 (80%) for the actual training. The overall number of active compounds in the independent dataset was 134, giving an active compound rate of 134/14,448=0.93%.

B. Selection of optimal descriptor set

A set of 1,252 descriptors in 35 categories was generated using the commercial ADRIANA.Code software [39]. The 35 categories consist of eight scalar descriptors, eight 2D and 3D auto-correlation functions each, eight radial distribution functions, and three surface-auto-correlation functions (see Table I).

Table I
The original molecular Descriptors by category

	Descriptor Name	Description
Scalar descriptors	Weight	Molecular weight of compound
	HDon	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule
	HAcc	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule
	XlogP	Octanol/water partition coefficient in [log units] of the molecule following the XlogP approach
	TPSA	Topological polar surface area in [Å^2] of the molecule derived from polar 2D fragments
	Polariz	Mean molecular polarizability in [Å^3] of the molecule
	Dipol	Dipole moment in [Debye] of the molecule
	LogS	Solubility of the molecule in water in [log units]
Vector descriptors 2D Autocorrelation (11 descriptors) / 3D Autocorrelation (12 descriptors) / Radial Distribution Function (128 descriptors)	Ident	weighted by atom identities
	SigChg	weighted by σ atom charges
	PiChg	weighted by π atom charges
	TotChg	weighted by sum of σ and π charges
	SigEN	weighted by σ atom electronegativities
	PiEN	weighted by π atom electronegativities
	LpEN	weighted by lone pair electronegativities
	Polariz	weighted by effective atom polarizabilities
Surface autocorrelation (12 descriptors)	ESP	Autocorrelation functions weighted by the molecular electrostatic potential
	HBP	Autocorrelation functions weighted by the hydrogen bonding potential
	HPP	Autocorrelation functions weighted by the hydrophobicity potential

To assess the capability of machine learning approaches for ‘classic’ QSAR, an ANN was trained only on the eight scalar descriptors. The quality of the model was measured by generating a ROC curve from the independent data set (Fig. 6). Implementing all 1,252 descriptors into the input gives a significant improvement in the quality of the model (Fig. 6).

In the next step systematically the least significant input parameters were removed to reduce noise. To determine the significance of each input the ANN is considered to be a multidimensional function

$$y = f(x_1, x_2, ..., x_n) \qquad (13)$$

with input values $x_1, \ldots, x_n$ and output y. Then the sensitivity of each input can be measured by determining the partial derivative

$$\frac{\partial^i f}{\partial x_i} \qquad (14)$$

The mean value of the sensitivity for each descriptor category determined which categories would be removed. E.g., sensitivities for all 128 inputs of a Radial Distribution Function were summed up and divided by 128. The categories with the lowest input sensitivity were sorted out. In a first step, the number of descriptors was reduced from 1,252 to 428 (Fig. 4). The quality of the trained ANN improved further. After a final pruning step 276 descriptors were kept for the final model. This optimized descriptor set was applied to all SVM and DT models to assure comparability.

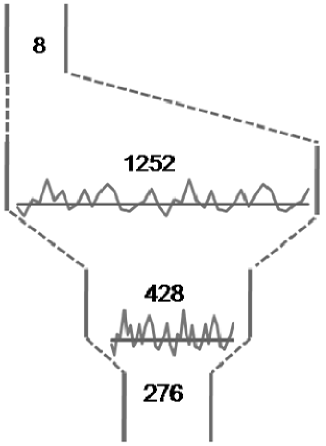

Fig. 4. The eight scalar descriptors were first expanded to all 1,252 by the addition of vector descriptors, and then sequentially reduced in size as the usefulness of each descriptor was determined.

C. Quality Measures of Machine Learning Methods

The machine learning approaches are evaluated by means of a receiver operating characteristic (ROC) curve. ROC curves plot the rate of true positives versus the rate of false positives, or sensitivity versus (1 - specificity). The diagonal represents performance of a random predictor. The larger the area under the curve (AUC) and the steeper the slope ascends towards the upper bound of the ROC curve (Fig. 5) the better the model.

The initial slope of the ROC curve relates to the “enrichment.” Enrichment is calculated by the following formula:

$$enrichment = \frac{\dfrac{TP}{TP+FP}}{\dfrac{P}{P+N}} \qquad (15)$$

where *TP* is the number of true positives, *FP* the number of false positives, *P* the total number of positives and *N* the total cases known to be negative. The value represents the factor by which the fraction of active compounds is increased in an *in silico* screened dataset.

V. Results

The ANNs were trained with up to 40,000 steps of Resilient Propagation. The training took 13 hours per network using in parallel eight cores of a Core 2 Quad 2.33GHz Intel Xeon microprocessor on the 64-bit version of Red Hat Enterprise Linux 5.2. The relative *rmsd* of the training data set reached 0.18, the monitor 0.22, and the independent data set 0.24.

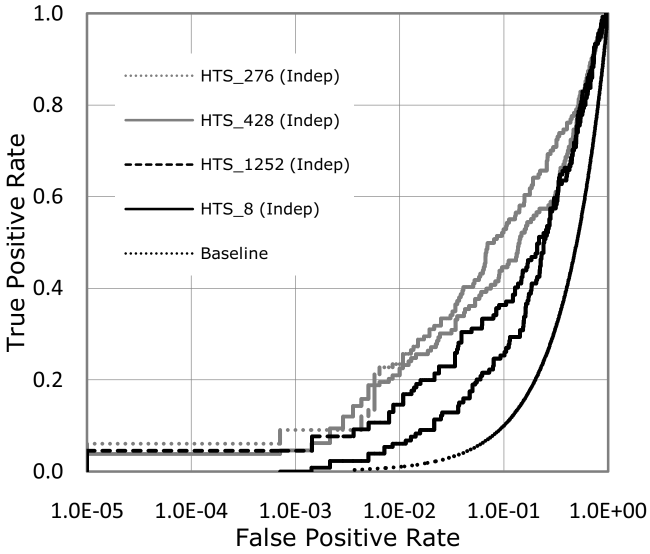

Fig. 5. Several ROC curves for the ANN. The black curve was achieved with a dataset with 8 descriptors, the dashed black line is based on data with 1,252 descriptors. The grey line represents the 428 descriptor dataset (all scalar, RDF_LpEN/PiEN/Polariz, 3DACorr_LpEN, SurfACorr_HBP/HPP), whereas the dotted gray line is the optimized dataset based on 276 descriptors (all scalar, RDF_LpEN/PiEN, 3DACorr_LpEN). The horizontal axis has a logarithmic scale.

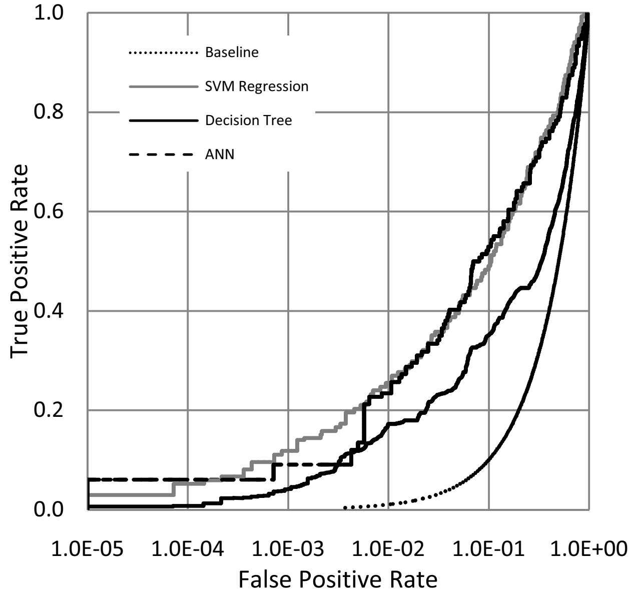

Fig. 6. ROC curves for all three methods on a logarithmic scale. The baseline is dotted, the solid black line is the DT, dashed line is the ANN, and the solid gray line is the SVM. Again, the scale of the horizontal axis is logarithmic.

TABLE II
AUC AND ENRICHMENT FOR ALL THREE METHODS

Method	ANN	SVM	DT
AUC	0.77	0.78	0.63
enrichment	61	64	43

To test the predictive capabilities of the ANN, it was applied to the independent dataset containing compounds not used in training. At a cutoff of 2μM, the ANN predicted 19 compounds to be active, eleven of which are known to be active (*TP*) and eight compounds to be inactive (*FP*), giving an enrichment of 61 (=11/19*144,475/1,356).

In a similar procedure the SVM was trained and the compounds in the independent data set were predicted for their activity. Considering just the top 10th percentile of the data at a cutoff of 7 μM, 20 molecules were predicted to be active with 12 of them being true positives. This results in an enrichment of 64 (=12/20*144,475/1,356) which is slightly better than the performance of the ANNs. The enrichments and AUCs achieved can be seen in Table II

FIRM was trained on the training dataset and the generated partitioning data were then applied to the independent dataset. Each compound was evaluated and assigned a predicted activity ranging from 0 to 100 for each compound, zero denoting a total lack of projected activity, and 100 signifying a fully active compound.

Each of the descriptor variables was evaluated individually, including the elements of the vector descriptors, and assigned merge/split values of 2.9/3.0. The maximum number of groups to be analyzed was set to 800, the minimum conservative significance level was given as 0.25, and all other parameters were left at their default values. These parameters were optimized on the monitoring dataset. Generation of the DT took slightly over an hour on a single core of a quad-core 3GHz Intel Xeon microprocessor under the 64-bit version of Red Hat Enterprise Linux 5.2.

Looking only at the top percentile, 34 compounds were predicted to be active, and 14 of them were true positives. The enrichment was therefore 43, which is slightly reduced when compared to the enrichment achieved by the other methods. Fig. 6 shows the ROC curves for the three methods.

To determine the potential benefit from applying a combination of these machine learning techniques the predicted compounds were analyzed for overlap. The results are given in the table below.

TABLE III
TRUE POSITIVES (TP) AND FALSE POSITIVES (FP) ACHIEVED BY EACH OF THE METHODS AND THEIR RESPECTIVE COMBINATIONS AS WELL AS THE RESULTING ENRICHMENT (E)

	ANN	SVM	DT	ANN SVM	ANN DT	DT SVM	All
TP	21	21	16	12	8	9	5
FP	94	36	54	5	6	5	3
E	19	39	24	75	61	68	67

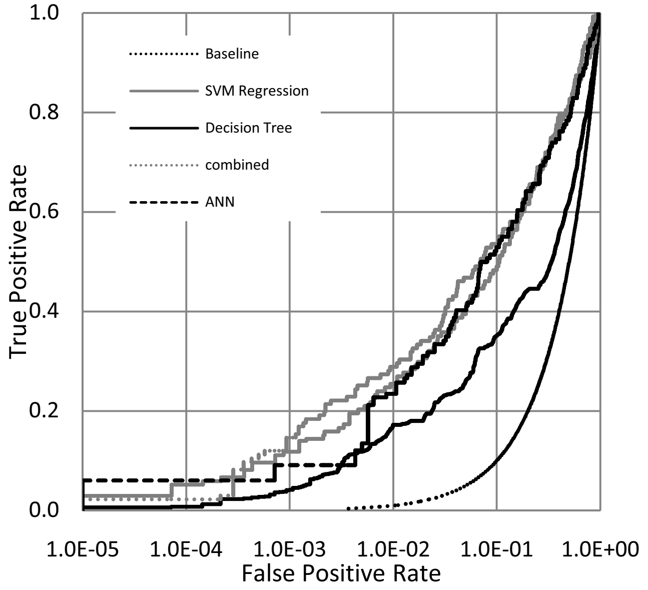

Fig. 7. This figure compares the ROC curves for the ANN, SVM, and DT (same line attributes as in Fig. 6) with a ROC curve created by averaging the values of the predictions made by the ANN and SVM (dotted gray line). The improvement is clear. Here too, the horizontal axis features a logarithmic scale.

The enrichment improves for all combinations. However, if all three methods are combined, only a small number of active compounds is recovered. Therefore, it is optimal to consider molecules designated as active by at least two of the methods. Fig. 7 shows an improved ROC curve (dotted gray line) obtained by combining the activity predictions made by the ANN and SVM, as well as the ROC curves of the individual methods for comparison.

VI. Conclusions

We present the application of a combination of machine learning techniques to drug discovery based on 3D QSAR models derived from HTS. All three tools achieve a high level of predictive accuracy in the process. The SVM, however, was best at predicting potentiators, whereas the DT approach produced the least accurate results. Furthermore, combining the SVM and ANN produced a visible improvement in overall predictive ability. Each of these methods, as well as their combinations, enable independent datasets to be enriched for active compounds by factors of 40-65 *in silico*. A problem-optimized descriptor set maximizes prediction accuracy for the presented project of finding new mGluR5 potentiators.

The methods showed similar performance on the independent data sets. Enrichment, ROC curve, and number of active and inactive compounds are in the same range. However, SVM performs slightly better than ANN, which has in turn an advantage over the DT. The highest enrichment is achieved by considering the compounds predicted active by both the ANN and SVM.

References

[1] J. Carballeira, M. Quezada, E. Alvarez, and J. Sinisterra, "High-throughput screening and QSAR-3D/CoMFA: useful tools to design predictive models of substrate specificity for biocatalysts," *Molecules,* vol. 9, pp. 673-693, 2004.

[2] H. Gao, M. S. Lajiness, and J. V. Drie, "Enhancement of binary QSAR analysis by a GA-based variable selection method," *Journal of Molecular Graphics and Modeling,* vol. 20, pp. 259-268, 2002.

[3] A. Böcker, G. Schneider, and A. Teckentrup, "Status of HTS data mining approaches," *QSAR & Combinatorial Science,* vol. 23, pp. 207-213, 2004.

[4] I. V. Tetko*, et al.*, "Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," *J. Chem. Inf. Model.,* vol. 48, pp. 1733-1746, 2008.

[5] D. Winkler, "Neural networks as robust tools in drug lead discovery and development," *Molecular Biotechnology,* vol. 27, pp. 139-167, 2004.

[6] W. W. Van Osdol, T. G. Myers, J. N. Weinstein, L. J. Michael, and B. Ludwig, "Neural network techniques for informatics of cancer drug discovery," in *Methods in Enzymology*. vol. 321, ed: Academic Press, 2000, pp. 369-395.

[7] A. Khandelwal*, et al.*, "Machine learning methods and docking for predicting human pregnane X receptor activation," *Chem. Res. Toxicol.,* vol. 21, pp. 1457-1467, 2008.

[8] T. Schroeter*, et al.*, "Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules," *Journal of Computer-Aided Molecular Design,* vol. 21, pp. 651-664, 2007.

[9] P. Tino, I. T. Nabney, B. S. Williams, J. Losel, and Y. Sun, "Nonlinear prediction of quantitative structure-activity relationships," *J. Chem. Inf. Comput. Sci.,* vol. 44, pp. 1647-1653, 2004.

[10] J. Burton*, et al.*, "Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset," *J. Med. Chem.,* vol. 49, pp. 6231-6240, 2006.

[11] D. Hecht, M. Cheung, and G. B. Fogel, "QSAR using evolved neural networks for the inhibition of mutant PfDHFR by pyrimethamine derivatives," *Biosystems,* vol. 92, pp. 10-15, 2008.

[12] D. Hecht and G. Fogel, "High-throughput ligand screening via preclustering and evolved neural networks," *IEEE/ACM Trans. Comput. Biol. Bioinformatics,* vol. 4, pp. 476-484, 2007.

[13] J. Fang, Y. Dong, G. H. Lushington, Q. Z. Ye, and G. I. Georg, "Support vector machines in HTS data mining: type I MetAPs inhibition study," *J Biomol Screen,* vol. 11, pp. 138-144, 2006.

[14] D. Plewczynski*, et al.*, "Target specific compound identification using a support vector machine," *Combinatorial Chemistry High Throughput Screening,* vol. 10, pp. 189-196, 2007.

[15] L. S. Florence and B. Jurgen, "Virtual screening methods that complement HTS," *Combinatorial Chemistry & High Throughput Screening,* vol. 7, pp. 259-269, 2004.

[16] Y. Sakiyama*, et al.*, "Predicting human liver microsomal stability with machine learning techniques," *Journal of Molecular Graphics and Modeling,* vol. 26, pp. 907-915, 2008.

[17] N. Baurin*, et al.*, "2D QSAR Consensus prediction for high-throughput virtual screening. an application to COX-2 inhibition modeling and screening of the NCI database," *J. Chem. Inf. Comput. Sci.,* vol. 44, pp. 276-285, 2004.

[18] A. Rusinko, M. W. Farmen, C. G. Lambert, P. L. Brown, and S. S. Young, "Analysis of a large structure/biological activity data set using recursive partitioning," *J. Chem. Inf. Comput. Sci.,* vol. 39, pp. 1017-1026, 1999.

[19] K. Simmons*, et al.*, "Practical outcomes of applying ensemble machine learning classifiers to high-throughput screening (HTS) data analysis and screening," *J. Chem. Inf. Model.,* vol. 48 pp. 2196–2206, 2008.

[20] K. Simmons*, et al.*, "Comparative study of machine-learning and chemometric tools for analysis of in-vivo high-throughput screening data," *J. Chem. Inf. Model.,* vol. 48, pp. 1663-1668, 2008.

[21] P. J. Conn, G. Battaglia, M. J. Marino, and F. Nicoletti, "Metabotropic glutamate receptors in the basal ganglia motor circuit," *Nat Rev Neurosci,* vol. 6, pp. 787-798, 2005.

[22] A. Ritzén, "Molecular pharmacology and therapeutic prospects of metabotropic glutamate receptor allosteric modulators," *Basic & Clinical Pharmacology & Toxicology,* vol. 97, pp. 202-213, 2005.

[23] X. H. Zhou, P. Castelluccio, and C. Zhou, "Nonparametric estimation of ROC curves in the

absence of a gold standard " *Biometrics,* vol. 61, pp. 600-609, 2005.

[24] J. Zupan and J. Gasteiger, *Neural networks for chemists : an introduction*. Weinheim ; New York: VCH, 1993.

[25] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *IEEE International Conference on Neural Networks,*, 1993, pp. 586-591.

[26] A. J. Smola and B. Schoelkopf, "A tutorial on support vector regression," vol. 14, ed: Kluwer Academic Publishers, 2004, pp. 199-222.

[27] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support vector regression machines* vol. 9. Cambridge, MA: MIT Press, 1997.

[28] B. Schoelkopf and A. J. Smola, *Learning with Kernels*. Cambridge, Massachusetts: The MIT Press, 2002.

[29] V. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1999.

[30] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics,* vol. 29, pp. 119-127, 1980.

[31] J. R. Quinlan, *C4.5: programs for machine learning*: Morgan Kaufmann Publishers Inc., 1993.

[32] R. Dingledine, K. Borges, D. Bowie, and S. F. Traynelis, "The glutamate receptor ion channels," *Pharmacol Rev,* vol. 51, pp. 7-62, 1999.

[33] P. J. Conn and J. P. Pin, "Pharmacology and functions of metabotropic glutamate receptors," *Annu Rev Pharmacol Toxicol,* vol. 37, pp. 205-37, 1997.

[34] J. P. Pin and R. Duvoisin, "The metabotropic glutamate receptors: structure and functions," *Neuropharmacology,* vol. 34, pp. 1-26, 1995.

[35] R. Anwyl, "Metabotropic glutamate receptors: electrophysiological properties and role in plasticity," *Brain Research Reviews,* vol. 29, pp. 83-120, 1999.

[36] V. Coutinho and T. Knopfel, "Metabotropic glutamate receptors: electrical and chemical signaling properties," *Neuroscientist,* vol. 8, pp. 551-561, 2002.

[37] M. J. Marino and P. J. Conn, "Modulation of the basal ganglia by metabotropic glutamate receptors: potential for novel therapeutics," *Current Drug Targets-CNS & Neurological Disorders,* vol. 1, pp. 239-250, 2002.

[38] B. Moghaddam, "Targeting metabotropic glutamate receptors for treatment of the cognitive symptoms of schizophrenia," *Psychopharmacology,* vol. 174, pp. 39-44, 2004.

[39] "Molecular Networks GmbH: Adriana.Code," *http://www.molecular-networks.com,* 2008.

[40] D. M. Hawkins, "FIRM: Formal Inference-Based Recursive Modeling," *The American Statistician,* vol. 45, p. 155, 1991.